ABSTRACT OF THE DISCLOSURE

A highly scalable system and method for supporting (mim,max) based Service Level Agreements (SLA) on outbound bandwidth usage for a plurality of customers whose applications (e.g.,Web sites) are hosted by a server farm that consists of a very large number of servers. The system employs a feedback system that enforces the outbound link bandwidth SLAs by regulating the inbound traffic to a server or server farm. Inbound traffic is admitted to servers using a rate denoted as Rt(i,j), which is the amount of the i^{th} customer's j^{th} type of traffic that can be admitted within a service cycle time to servers which support the i^{th} customer. A centralized device computes Rt(i,j) based on the history of admitted inbound traffic to servers, the history of generated outbound traffic from servers, and the SLAs of various customers. The Rt(i,j) value is then relayed to one or more inbound traffic limiters that regulate the inbound traffic using the rates Rt(i,j) in a given service cycle time. The process of computing and deploying Rt(i,j) values is repeated periodically. In this manner, the system provides a method by which differentiated services can be provided to various types of traffic, the generation of output from a server or a server farm is avoided if that output cannot be delivered to end users, and revenue can be maximized when allocating bandwidth beyond the minimums.